Lightening the Load of Document Smoothing for Better Language Modeling Retrieval

Mark D. Smucker and James Allan
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts Amherst

ABSTRACT

We hypothesized that language modeling retrieval would improve if we reduced the need for document smoothing to provide an inverse document frequency (IDF) like effect. We created inverse collection frequency (ICF) weighted query models as a tool to partially separate the IDF-like role from document smoothing. Compared to maximum likelihood estimated (MLE) queries, the ICF weighted queries achieved a 6.4% improvement in mean average precision on description queries. The ICF weighted queries performed better with less document smoothing than that required by MLE queries. Language modeling retrieval may benefit from a means to separately incorporate an IDF-like behavior outside of document smoothing.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Algorithms, Experimentation

Keywords: Inverse Document Frequency, IDF, Document Smoothing, Language Modeling

1. INTRODUCTION

The language modeling approach to information retrieval represents queries and documents as probabilistic models [1]. While inverse document frequency (IDF) is common to many retrieval methods, it is not explicitly a part of the language modeling approach to retrieval. Zhai and Lafferty showed that language modeling contains an IDF-like like component when documents are smoothed with the collection model [2]. The IDF-like behavior is provided in the form of the inverse collection frequency (ICF). Zhai and Lafferty developed two-stage smoothing to separately control and leverage smoothing's *estimation* and *query modeling* capabilities [3]. Two-stage smoothing also eliminates the need for parameter tuning with training data. The query modeling role of smoothing describes the need for more smoothing to increase the IDF-like effect for verbose queries. Two-stage smoothing is competitive with, and sometimes outperforms, the best performance obtainable from Dirichlet prior and Jelinek-Mercer smoothing [3].

We hypothesize that providing the inverse collection frequency outside of document smoothing will result in improved retrieval. In a sense, removing a role from document smoothing will allow smoothing to perform better in its remaining roles.

To test our hypothesis, we weight our query models in proportion to the ICF. While this is an ad-hoc method to determine the probabilities of a query model, it does allow for the ICF to be partially separated from document smoothing.

2. METHODS

Maximum likelihood estimation (MLE) is a common technique for estimating query models. MLE estimates the probability of a word given a text as the count of that word divided by the total number of words in the text. As such, the MLE probability of a word w given a query Q is:

$$P(w|M_Q) = \frac{Q(w)}{|Q|}$$

where $Q(w)$ is the count of word w in the query Q and $|Q| = \sum_w Q(w)$ is the query's length.

We created our inverse collection frequency weighted query models as follows:

$$P(w|M_Q) = -\frac{1}{Z} Q(w) \log P(w|C) \qquad (1)$$

where $P(w|C)$ is the MLE model of the collection, and Z is the normalization factor so that the query model sums to 1. Taking the negative of the log, $-\log P(w|C)$, gives an amount proportional to the inverse collection frequency.

This ICF weighting approach is analogous to the inverse document frequency and may seem ad-hoc in construction. Our goal with the ICF weighted query is to provide a means to test our hypothesis. Our goal is not to propose a new way to estimate a query model. We hope that further work might result in a compelling formal integration of the idea into language modeling retrieval.

3. EXPERIMENTS

We compared the inverse collection frequency query models created via equation 1 to the MLE query models. For our queries, we used the title and description fields of the TREC topics 301-450, which are the ad-hoc topics for TREC

SIGIR'06, August 6–11, 2006, Seattle, Washington, USA.
ACM 1-59593-369-7/06/0008.

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **2006**	2. REPORT TYPE	3. DATES COVERED **00-00-2006 to 00-00-2006**

4. TITLE AND SUBTITLE
Lightening the Load of Document Smoothing for Better Language Modeling Retrieval

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Massachusetts Amherst,Center for Intelligent Information Retrieval,Department of Computer Science,Amherst,MA,01003

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**2**	

Title Queries

Test Set	MLE	ICF	% Imp.	Sig. Level
TREC 6	0.228	0.234	2.6%	0.06
TREC 7	0.182	0.189	3.9%	0.12
TREC 8	0.247	0.246	-0.3%	0.81
All: 6,7,8	0.219	0.223	1.9%	0.04

Description Queries

Test Set	MLE	ICF	% Imp.	Sig. Level
TREC 6	0.195	0.212	8.9%	< 0.01
TREC 7	0.183	0.194	6.4%	0.03
TREC 8	0.228	0.238	4.4%	0.05
All: 6,7,8	0.202	0.215	6.4%	< 0.01

Table 1: Mean average precision results for title and description queries. MLE is the maximum likelihood query. ICF is the query model constructed using the inverse collection frequency as per equation 1.

6, 7, and 8. The collection was TREC volumes 4 and 5 minus the Congressional Record. We used the Lemur retrieval toolkit for our experiments. We stemmed all words with the Krovetz stemmer and used an in-house stopword list of 418 words. We ranked documents by their cross entropy with the query.

We used three-fold cross validation to set the document smoothing parameters. Each set of TREC topics was a fold. For example, the TREC 7 and 8 topics were the training set for the TREC 6 topics. Thus our results closer represent expected performance than optimal. We did parameter sweeps of Dirichlet prior and Jelinek-Mercer smoothing. In all cases Dirichlet prior performed better. Dirichlet prior smoothing mixes the MLE document model with the MLE collection model: $P(w|M_D) = (1 - \lambda)P(w|D) + \lambda P(w|C)$, where $\lambda = m/(|D| + m)$ and m is the Dirichlet prior parameter.

We measured statistical significance with a paired, two-sided, randomization test with $100,000$ samples.

4. RESULTS AND DISCUSSION

Table 1 shows the mean average precision of MLE and ICF query models for title and description queries. ICF obtained a 1.9% improvement over MLE on title queries and a 6.4% improvement on description queries. For title queries, none of the individual TREC topic sets showed statistically significant improvements. For description queries, statistically significant improvements occurred for TREC 6 and 7. ICF's performance gain comes from improving precision at recall points of 0.3 and higher.

Table 2 shows that less document smoothing is required if the retrieval method has available another way to utilize the inverse collection frequency in an IDF-like manner.

Our results show that improved retrieval performance is possible if the IDF-like role played by document smoothing is separately handled. The performance improvement may be the result of either the better use of the ICF or better document smoothing or both.

Two-stage smoothing performs well, but in the configuration tested by Zhai and Lafferty, it does not result in statistically significant performance improvements on title or description queries for TREC 7 and 8 as compared to the best performance of Dirichlet prior or Jelinek-Mercer smoothing [3]. Two-stage smoothing separately models the estimation and query-modeling (IDF-like) roles of smoothing, but it still ties the IDF-like role to the amount of smoothing required. In contrast, when we separated the IDF-like role from smoothing, we obtained a modest, but statistically significant, performance improvement for description queries on TREC 7. While two-stage smoothing requires no parameter tuning, one could tune its parameters and possibly achieve similar improvements to those reported here.

	Dirichlet m	
Training Set	MLE	ICF
Titles TREC 6, 7	800	500
Titles TREC 6, 8	350	400
Titles TREC 7, 8	800	500
Desc. TREC 6, 7	2500	1250
Desc. TREC 6, 8	2500	1000
Desc. TREC 7, 8	3000	1250

Table 2: The best parameter setting for Dirichlet prior smoothing on the training sets. Except for the TREC 6, 8 titles, the inverse collection frequency (ICF) query models performed better with less smoothing than the maximum likelihood estimated (MLE) models.

5. CONCLUSION

Document smoothing plays several roles in language modeling retrieval [2]. One of these roles is to produce an IDF-like effect using the inverse collection frequency (ICF). We hypothesized that lightening the IDF-like responsibilities of smoothing could improve retrieval. We created ICF weighted query models and compared their performance to maximum likelihood estimated models. The ICF query models achieved a 6.4% improvement in the mean average precision while requiring less document smoothing. Language modeling retrieval may benefit from a means to incorporate an IDF-like behavior outside of document smoothing and thereby allow document smoothing to focus on its other roles.

6. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by the Defense Advanced Research Projects Agency (DARPA) under contract number HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

7. REFERENCES

[1] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR*, pages 275–281, 1998.

[2] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR*, pages 334–342, 2001.

[3] C. Zhai and J. Lafferty. Two-stage language models for information retrieval. In *SIGIR*, pages 49–56, 2002.